A data-driven approach to alternations based on protein-protein interactions

Schneider, Gerold ; Rinaldi, Fabio

Abstract: Syntactic alternations like the dative shift are well researched. But most decisions which speakers take are more complex than binary choices. Multifactorial lexicogrammatical approaches and a large inventory of syntactic patterns are needed to supplement current approaches. We use the term semantic alternation for the many ways in which a relation between entities, conveying broadly the same meaning, can be expressed. We use a well-resourced domain, biomedical research texts, for a corpusdriven approach. As entities we use proteins, and as relations we use interactions between them, using Text Mining training data. We discuss three approaches: first, manually designed syntactic patterns, second a corpus-based semi-automatic approach and third a machine-learning language model. The machine-learning approach learns the probability that a syntactic configuration expresses a relevant interaction from an annotated corpus. The inventory of configurations define the envelope of variation and its multitude of forms.

Originally published at:
Schneider, Gerold; Rinaldi, Fabio (2011). A data-driven approach to alternations based on protein-protein interactions. In: III Congreso Internacional de Lingüística de Corpus, Valencia, Spain, 7 April 2011 - 9 April 2011. Universitat Politècnica de València, 597-607.

A data-driven approach to alternations based on protein-protein interactions

GEROLD SCHNEIDER AND FABIO RINALDI

gschneid@ifi.uzh.ch[1] , rinaldi@ifi.uzh.ch

Institute of Computational Linguistics, University of Zurich

Syntactic alternations like the dative shift are well researched. But most decisions which speakers take are more complex than binary choices. Multifactorial lexicogrammatical approaches and a large inventory of syntactic patterns are needed to supplement current approaches. We use the term semantic alternation for the many ways in which a relation between entities, conveying broadly the same meaning, can be expressed. We use a well-resourced domain, biomedical research texts, for a corpus-driven approach. As entities we use proteins, and as relations we use interactions between them, using Text Mining training data. We discuss three approaches: first, manually designed syntactic patterns, second a corpus-based semi-automatic approach and third a machine-learning language model. The machine-learning approach learns the probability that a syntactic configuration expresses a relevant interaction from an annotated corpus. The inventory of configurations define the envelope of variation and its multitude of forms.

Keywords: syntactic alternations, lexicogrammar, corpus-driven, semantic alternation, text mining, machine learning

Alternaciones sintácticas como la alternancia de dativo se han investigado extensivamente. Sin embargo la mayoría de las decisiones que toman los hablantes van más alla de simples opciones binarias. Métodos multifactorial léxico-gramaticales y un amplio inventario de patrones sintácticos son necesarios para complementar los métodos actuales. Utilizamos el término alternancia semántica para indicar las distintas maneras de expresar una relación entre entidades con el mismo significado. Para nuestro estudio utilizamos como corpus artículos científicos del campo biomédico. Las entidades que consideramos son proteínas, genes, enfermedades y medicinas, y estudiamos las relaciones entres ellas. En nuestro artículo presentamos tres métodos: en primer lugar, patrones sintácticos desarollados manualmente, en segundo lugar un enfoque semi-automático basado en corpus y tercero un enfoque que utiliza técnicas de Aprendizaje Automático. El sistema de Aprendizaje Automático extrae de un corpus anotado la probabilidad que una configuración sintáctica específica exprese una interacción relevante. El inventario de las configuraciones permite definir las variaciones sintácticas en todas sus formas.

Palabras clave: Alternaciones sintácticas, léxico-gramática, lingüística de corpus, alternancia semántica, text mining, aprendizaje automático

[1] Corresponding author

1 INTRODUCTION

The present paper suggests the use of a corpus-driven approach to alternations.[2] Instead of viewing alternations as a binary decision between two choices we suggest a view of alternations as a multifactorial phenomenon of many choices, relating the many different ways of expressing similar concepts to each other. We use a *corpus-driven* approach. Instead of focussing on a single phenomenon and its *envelope of variation* (Labov, 1969), a corpus-driven approach forces the researcher to interpret many features, which possibly interact with each other. The detection of the envelope of variation is typically not corpus-driven and often not clear. For example, Arppe, Gilquin, Glynn, Hilpert and Zeschel (2011) state:

> Our focus on alternations is the result of theoretical heritage from generative syntax and a matter of methodological convenience. Most linguistic decisions that speakers make are more complex than binary choices … alternations are as simplistic and reductionistic as the theories of language that originally studied them (Arppe *et al.,* 2011).

Corpus-driven approaches are for example used to discover collocations (Evert, 2008), or diachronic word class shifts (Mair, Hundt, Leech & Smith, 2002). For the discovery of collocations, word forms or lemmas are used as uncontested features, for word-class shifts agreed-on part-of-speech tags can be used. In the case of alternations, there is a considerably less stable base than in collocations or part-of-speech tags, as Arppe *et al.* (2011) warn us. In particular, there are manifold restrictions, strong lexicogrammatical interactions, the sheer number of alternation is contested. In other words: in (probably) the majority of cases where an alternation could syntactically be used, it will lead to a different semantic or an inacceptable or at least not native-like utterance (Pawley & Syder, 1983). In computational terms, there is a precision problem: many application of an alternation rule lead to incorrect results. Also, the vast majority of utterances where two speakers express the same concept differs in more respects than in the choice of a single alternation.

As *semantic equality* is the touchstone of alternation (only those applications of an alternation rule that keep the semantic content largely unchanged are part of the envelope of variation), we would like to keep it as a base. Classical approaches to alternations start from the precision perspective: apply alternations and overgenerate; lose on recall

[2] This research is partially funded by the Swiss National Science Foundation (grant 100014-118396/1). Additional support is provided by Novartis Pharma AG, NITAS, Text Mining Services, CH-4002, Basel, Switzerland.

anyway. We would like to suggest starting from a recall perspective: aim at collecting and recognizing all utterances that express the same concept and find out which complex set of alternation choices were involved.

Although we base our suggestions on large amounts of data and carefully annotated corpora, our investigation is exploratory in nature. In section 2, we give a brief introduction to the concept of corpus-drivenness. In section 3 we illustrate and motivate our view of alternations as a multifactorial phenomenon. In section 4, we present our method for collecting and detecting different ways of expressing the same concept, based on carefully annotated corpora from carefully restricted concepts, using an Information Retrieval approach.

2 THE CORPUS-DRIVEN APPROACH

The distinction between *corpus-driven* and *corpus-based* has been described by Tognini-Bonelli (2001). In corpus-based approaches, existing hypothesis are tested, while in corpus-driven or *data-driven* approaches, hypotheses arise from the corpus data. Corpus-driven approaches have a advantages and disadvantages. An advantage is that, in areas of gradience and subtle differences, it can bring patterns to the surface that went unnoticed by linguists (e.g. Hunston & Francis, 2000). Variationist linguistics is often very subtle and gradient.

A disadvantage of corpus-driven approaches is that they rely on the quality of the corpus: "… since the information provided by the corpus is placed centrally and accounted for exhaustively, then there is a risk of error if the corpus turns out to be unrepresentative" (Tognini-Bonelli, 2001:88). For corpus-driven approaches, large amounts of data are necessary, and relying on frequencies implies a tacit hypothesis, namely that significant frequency differences in the investigated data are indicative.

3 ALTERNATIONS AS AN OPEN SET

In a classical approach, alternations are two syntactic configurations that are used to convey the same meaning. In English, well-known examples are the dative shift which

links sentences (1) and (2), or the Genitive alternation, which links sentences (3) and (4).

(1) Peter gave Mary a book.

(2) Peter gave a book to Mary.

(3) Mary's car is fast.

(4) The car of Mary is fast.

3.1 Restrictions on alternations

There are many restrictions on the application of alternations. These restrictions are typically referred to as the *envelope of variation* (Labov, 1969) or the *choice context* (Rosenbach, 2003). These restrictions rule out contexts in which a speaker does not really have a choice between the two variants. For example, while

(5) Peter gave a book to those students who had achieved a grade A mark.

is acceptable, hardly anybody would produce (6):

(6) ?Peter gave those students who had achieved a grade A mark a book.

(6) violates the linguistic tendency to put long constituents after short ones, the principle of end weight. Similarly, while

(7) Mary's picture of the house is great.

is acceptable, (8) is highly unusual, because nested Saxon genitives are extremely rare, and because the of-PP in (7) does not necessarily express a possessor relation.

(8) ?Mary's house's picture is great.

There are at least a dozen such restrictions for each alternation (e.g. Jucker, 1993). Only a minority of all candidate configuration tokens are really available for the alternation. While syntactic restrictions can be listed, there is an almost infinite set of semantic restrictions. The verbal semantics of *give*, for example entail that sentences (9) and (10) are only equivalent if the printout of a speech is intended.

(9) Mary gave a speech to the students.

(10) Mary gave the students a speech.

The deep-syntactic role typically depends on verb semantics. In the nominalization alternation, for example, *destruction of the city* implies *city* as object, while in *implication of the discovery* the word *discovery* is a subject. Such behaviour can be found in most alternations. For example, *God's* creation and *the creation of God* are probably not in the envelope of variation.

3.2 Interactions between Lexis and Grammar

Interactions between lexis and grammar have been investigated for the dative shift, e.g. Bresnan and Nikitina (2009) or Lehmann and Schneider (2011). Pronouns as indirect object favour the double NP construction, and there are many idioms and collocational preferences. For example, in *give birth to baby* the alternative *give a baby birth* is basically not used. The configurations favouring the double NP-construction most strongly in Lehmann and Schneider (2011) are given in table 1.

lemma triplet	dshift	to	for	% dshift	iObj
ask you question	4876	3	8	99.8	you
tell you truth	1203	4	1	99.6	you
tell you story	958	3	3	99.4	you
ask him question	1089	6	1	99.4	him
show you picture	1698	13	1	99.2	you
give you number	470	3	1	99.2	you
bring you update	456	5	0	98.9	you
give them information	519	6	0	98.9	them
bring them home	502	6	0	98.8	them
ask them question	404	3	2	98.8	them

Table 1. Dative shift lemma triplets ordered by preference for the double-NP construction

Non-core ditransitive verbs have a different behaviour from prototypes. The most prototypical verb, *give*, has a preference for the double-NP construction, while marginal ditransitives such as *provide*, are rarely used with the double-NP construction.

It is unclear if a list of ditransitive verbs can be compiled in the first place. There are indications that they form an open class. Lehmann and Schneider (2011), for example, deliver the following examples.

(11) One husband, accompanying his wife to a fitting there, responded to her lament that she had nowhere to wear the ballgown he had selected for her by promising to **throw** the ***dress*** a lavish ***party***. (TLN956252198)

(12) **Cry** ***Orwell*** a ***river***, Mr. Timberlake, while CNN's Jeanne Moos reminds us what 20 years of VMAs have really been all about. (CNN:20030827SE.02)

(13) By moving to pastures new, successful managers can **negotiate** ***themselves*** a new ***package*** of options from scratch. (BNC:AH2:375)

3.3 Alternations are not a binary switch

The verb *provide* which we have mentioned also illustrates that the alternation can take many forms: the double-NP construction is not the alternative to an NP + of-PP construction, but to an NP + with-PP construction, or an NP + to-PP construction. The double-NP construction is markedly rare, the BNC contains a few dozen of double-NP constructions, about 4000 with-PP constructions, and about 2000 to-PPs. Examples are given in (14) to (16), the automatic syntactic dependency analysis of sentence (14) is also given in figure 1.

(14) You provided him his death, others have provided him a grave. (BNC-Wri K8S)

(15) The forwards played extremely well as a unit, driving in unison and **providing** their backs **with good ball**. (BNC-Wri K5A)

(16) Salespeople may also be called upon **to provide** after-sales service **to customers**. (BNC-Wri K94)

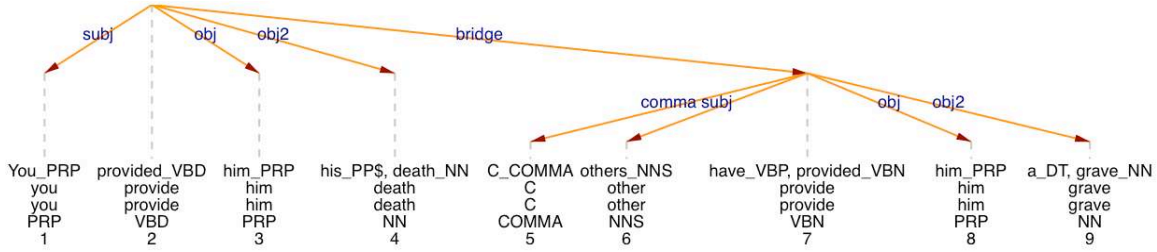

Figure 1. Syntactic analysis of sentence (14).

Many verbs also have an NP + for-PP alternative, which typically expresses benefactor. Levin (1993) provides a detailed verb-lexicon based analysis of alternations. Based on the observation that *load X onto Y* (e.g. sentence (17)) and *load Y with X* (e.g. sentence (18)) express the same meaning she created alternation classes.

(17) In June 1989 the East Londoner had taken a load to Barcelona, where depot staff **loaded his trailer with a mixed consignment** to be taken back to London. (BNC-Wri AHM)

(18) The men, from Pickfords Removals, **were loading a machine onto a trailer** when part of it collapsed, trapping the men beneath them. (BNC-Wri K1G).

Levin compiled 51 coarse classes, containing a total of 193 fin-grained classes. Levin classes cover 3100 verbs.

Ditransitive constructions, in fact most constructions, need to be disambiguated in the context, which means that a dictionary-based approach like Levin's will massively overgenerate, and token-wise disambiguation is necessary.

(19) But I **call** you **a whore**! (BNC-Wri FEE)

(20) Shall I get them **to call** you **a cab**? (BNC-Wri G0B)

PropBank (Palmer, Gildea, & Kingsbury, 2005) and FrameNet (Fillmore, Johnson, & Petruck, 2003) are projects that assign *thematic roles* to verbs and their arguments in context (see Baker and Ruppenhofer (2002) for a comparison between Levin and FrameNet), and that have been used for token-wise disambiguation, for example in the CoNLL-2005 shared task (Carreras & Màrquez 2005). The baseline performance was about 40% F-Score. A baseline takes simple class-based decisions, for example assigning the thematic role *agent* to all subjects and *patient* to all objects. The best system described in Carreras and Màrquez (2005), Punyakanok, Koomen, Roth and Yih (2005), reaches an F-Score of 79.44%. These two percentages illustrate the difficulty of the task – more difficult than for example syntactic parsing. Automatic syntactic parses were provided to the participants.

Lexis is a major disambiguation tool in such tasks, as far as the sparseness problem allows. For example, also humans largely disambiguate (19) and (20) based on the lexis of the second NP, (17) and (18) share the lexis of one object NP. Data sparseness is a problem as most lexical items are very rare (Zipf's law). Due to data sparseness, decisions of classifiers are usually taken at a level between the class-based baseline and a fully lexical decision.

3.4 Semantic Alternations

As semantic equality is the touchstone of alternation (only those applications of an alternation rule that keep the semantic content largely unchanged are part of the envelope of variation), we would like to keep it as a base. Classical approaches to alternations start from the *precision*-centered perspective: apply and overgenerate, filter with constraints; lose on recall anyway. We would like to suggest starting from a *recall*-centered perspective: aim at collecting and recognizing all utterances that express the same concept and find out which complex set of alternation choices were involved. Information Retrieval, in particular *Text Mining*, is an applied science that aims to find all textual forms which express a sought-for concept. The detection of *events* (also often termed *relations*) is particularly relevant for the domain of alternations. Events are typically verb-based, the participants of an event are arguments of the verb, and all configurations that are used to connect them to the verb should be detected.

4 METHODS

We use the following Text Mining scenario for our method: detection of *protein-protein* or *gene-disease-drug interactions* from biomedical texts. Biomedical Text Mining is a domain that has highly developed linguistic resources, for example protein databases, corpora that are annotated for events (IntAct, etc, REFs) and frequent shared tasks where state-of-the-art approaches are competing. In order to recognize a verb's syntactic arguments, we use a syntactic dependency parser (Schneider, 2008). We have used different approaches, which are briefly summarised in the following.

4.1 Manual Alternation Patterns

Initially we used a model freely combining classical alternations such as passive, dative shift, genitive, and nominalization. Although it was fairly successful, it overgenerated considerably (Rinaldi, Schneider, Kaljurand, Hess & Romacker, 2006).

4.2 Manual class-based disambiguation

Every sentence that contains at least two proteins can express a protein-protein relation. The manual annotation of the corpus, as well as the application phase, needs to discern between those syntactic connections that express a relevant interaction and those that do not. We use the following method: we collect all protein-pairs connected by a dependency chain ('path') from a term-annotated corpus (we use the GENIA corpus (Kim, Ohta, Tateisi & Tsujii, 2003)). We refer to the syntactic chain up from both proteins to where they meet as *path,* which is then used as a training *feature*. In figure 2, we see the path connecting the gene *nAChR* to the disease *schizophrenia*.

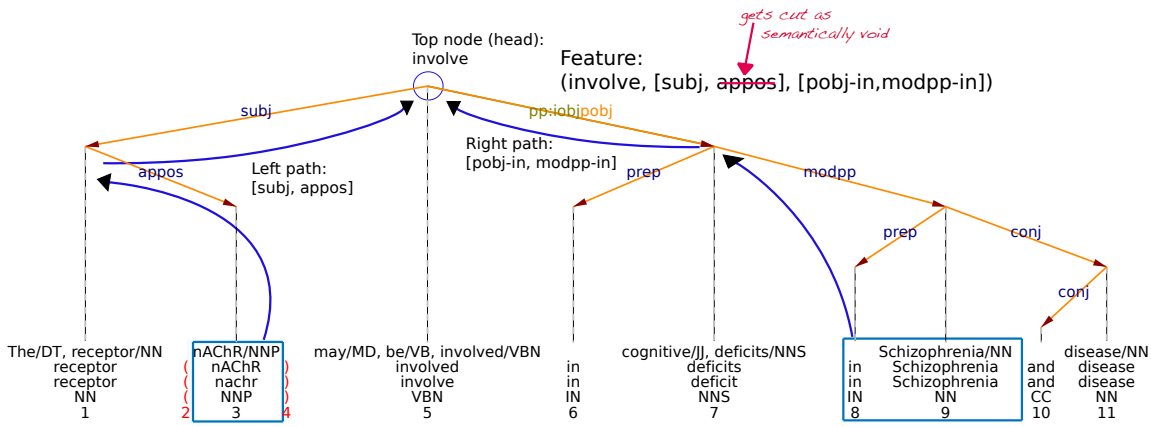

Figure 2. Syntactic path connecting the gene *nAChR* to the disease *schizophrenia*

Syntactic relations that are semantically void, like apposition and conjunction, are cut. The manual annotation decisions on which paths are relevant can be used directly for the application phase, augmented by a backoff chain to fight sparse data. The approach is described in detail in Schneider, Kaljurand and Rinaldi (2009). We have participated in the BioCreative II.5 competitive evaluation of biomedical text mining systems (Leitner, Mardis, Krallinger, Cesareni, Hirschmann & Valencia 2010). We achieved the best run for the detection of protein-protein interactions (according to the AUC iP/R metric (Manning, Raghavan & Schütze, 2008). Our system was overall considered as one of the best three.

4.3 Probabilistic Model

While the approach described in section 4.2 still involves a manual step, fully automatic learning is possible. If a sufficiently large training corpus is provided, the probability of being relevant for each syntactic path connecting any two entities (E1 and E2) can be calculated, using *Bayesian statistics*, as the following division: how often a given path expresses an interaction between the two entities (*count* in table 2), divided by how often the path appears between the two entities in the whole corpus (*potential* in table 2). We call a path relevant if it expresses an event, i.e. an interaction between two entities.

$$\mathrm{p}(\text{relevant} \mid \mathrm{path}_{E1,E2}) = \frac{f(\text{relevant}(\mathrm{path}_{E1,E2}))}{f(\mathrm{path}_{E1,E2})}$$

For each candidate entity pair, e.g. two proteins appearing in the same sentence, we suggest paths which have a probability above a certain threshold as being relevant. Many backoffs are used against sparse data. We have applied this approach e.g. in BioNLP 2009, and obtained good results (Kaljurand, Schneider & Rinaldi, 2009). Table 2 shows the most frequent counts from a training corpus for gene-disease-drug interactions.

Probability	Head	Path1	Path2	Count	Potential
13.62%	associate	subj	pobj-with	53	389
17.82%	associate	subj modpp-in	pobj-with	31	174
18.29%	cancer			30	164
14.57%	effect	modpp-of	modpp-on	22	151
18.92%	effect	modpp-of	modpp-on modpp-of	21	111
20.65%	association	modpp-of	modpp-with	19	92
6.29%	be	obj modpp-of	subj	19	302
17.82%	metabolize	pobj-by	subj	18	101
29.63%	inhibit	pobj-by	subj	16	54
35.71%	associate	subj modpp-in	pobj-with modpp-of	15	42

23.81%	cause	subj modpp-in	obj	15	63
5.02%	be	subj	obj modpp-of pobj-in modpart	15	299
100.00%	analyze	subj modpp-in	pobj-with	14	14

Table 2. Most frequent relevant paths between entities from a training corpus and probabilities of being relevant

We have also used versions in which the events are typed, thus forming semantic equivalence classes. All events of the same equivalence class can be said to be semantic alternations. Protein-protein interactions are for example often typed into classes like *regulation*, *binding* and *expression*. Although these classes are domain-specific, they have allowed us to construct a repository of semantic alternations for one domain, as a proof of concept for our suggested model of semantic alternations. There is a strong correlation between the head lexeme and the event type. Table 3 lists the four paths that are found in a training corpus for the verb (first 2 rows) and noun (last 2 rows) *influence*. The paths define the envelope of variation. The last two rows coincide with the classical passive alternation. The probability in the first column is straightforward to interpret in an Information Retrieval setting, but in our setting of finding alternations it is not clear how low a probability should be before we reject it a part of a variation envelope.

Probability	Head	Path1	Path2	Count	Potential
37.50%	influence	modpp-of	modpp-on modpp-of	9	24
21.88%	influence	modpp-of	modpp-on	7	32
5.88%	influence	subj	obj	6	102
44.44%	influence	subj modpp-of	pobj-by	4	9

Table 3. Data-driven alternations of *influence*

5 Conclusions

We have suggested a model of semantic alternations which does not use the classical precision-centered perspective (apply and overgenerate, filter with constraints; lose on recall anyway) but an approach starting from a recall-centered perspective: aim at collecting and recognizing all utterances that express the same concept and find out which complex set of alternation choices were involved. We have presented an Information Retrieval application for the biomedical genre which implements this perspective, and which has delivered good results. We use this linguistic model for semantic alternations as a proof of concept.

6 REFERENCES

Arppe, A., Gilquin, G., Glynn, D., Hilpert, M., & Zeschel, A. (2011). Cognitive Corpus Linguistics: Five points of debate on current theory and methodology. To appear in *Corpora* 5/2.

Baker, Collin F. & Ruppenhofer, J. (2002) FrameNet's Frames vs. Levin's Verb Classes. In J. Larson and M. Paster (Eds.) *Proceedings of the 28th Annual Meeting of the Berkeley Linguistics Society*. 27-38.

Bresnan, Joan & Nikitina, T. (2009). The Gradience of the Dative Alternation. In L. Uyechi & L. Hee Wee (Eds.), *Reality Exploration and Discovery: Pattern Interaction in Language and Life*. Stanford: CSLI Publications.161-184.

Carreras, X. & Màrquez, L. (2005). Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling. *In Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, Ann Arbor, Michigan, 152-164.

Evert, S. (2008). Corpora and collocations. In A. Lüdeling & M. Kytö (Eds.), Corpus Linguistics. An International Handbook, article 58. Berlin: Mouton de Gruyter.

Fillmore, C. J., Johnson, C. R. & Petruck, M. R. L. (2003). Background to FrameNet. *International Journal of Lexicography*, 16:235–250.

Hunston, S. & Francis, G. (2000). Pattern grammar: A corpus-driven approach to the lexical grammar of English. Amsterdam: Benjamins,

Jucker, A. H. (1993). The Genitive versus the of-Construction in Newspaper Language. In: A.H. Jucker (Ed). *The Noun Phrase in English: Its Structure and Variability*. Heidelberg: Universitätsverlag Winter. 121-136.

Kaljurand, K, Schneider, G. & Rinaldi, F. (2009). UZurich in the BioNLP 2009 Shared Task. In *Proceedings of BioNLP workshop*, NAACL/HLT, Boulder, Colorado.

Kim, J., Ohta, T., Tateisi, Y. & Tsujii, J. (2003). GENIA corpus - a semantically annotated corpus for bio-textmining, *Bioinformatics* 19 (1), 180-182.

Labov, W. (1969). Contraction, deletion, and inherent variability of the English copula *Language 45. 4. 715-62.*

Lehmann, H. M. & Schneider, G. (2011). Syntactic variation and lexical Preference in the Dative Shift Alternation. Paper presented at ICAME 2010. To appear in *VariEng*.

Leitner, F., Mardis, S. A., Krallinger, M., Cesareni, G., Hirschman, L. A. & Valencia, A. (2010). An Overview of BioCreative II.5. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 385-399.

Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation.* Chicago: University of Chicago Press.

Mair, C., Hundt, M., Leech, G., & Smith, N. (2002). Short term diachronic shifts in part-of-speech frequencies. A comparison of the tagged LOB and F-LOB corpora. *International Journal of Corpus Linguistics*, 7(2), 245-264.

Manning, C.D., Raghavan, P. & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Palmer, M., Gildea, D. & Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1).

Pawley, A. & Syder, F. H. (1983). Two Puzzles for Linguistic Theory: Native-like selection and native-like fluency. In J. C. Richards, and R. W. Schmidt (Eds.), *Language and Communication.* London: Longman. 191–226.

Punyakanok, V., Koomen, P., Roth, D., & tau Yih, W. (2005). Generalized inference with multiple semantic role labeling systems. In *Proceedings of CoNLL-2005*.

Rinaldi, F., Schneider, G., Kaljurand, K., Hess, M. & Romacker, M. (2006). An environment for relation mining over richly annotated corpora: the case of GENIA. *BMC Bioinformatics*, 7(Suppl. 3).

Rosenbach, A. (2003). Aspects of iconicity and economy in the choice between the s-genitive and the of-genitive in English. In G. Rohdenburg & B. Mondorf (Eds). *Determinants of Grammatical Variation in English*. Berlin/New York: Mouton de Gruyter. 379-411.

Schneider, G. (2008). *Hybrid Long-Distance Functional Dependency Parsing*. Doctoral Thesis. Institute of Computational Linguistics, University of Zürich.

Schneider, G., Kaljurand, K. & Rinaldi, F. (2009). Detecting Protein-Protein Interactions in Biomedical Texts using a Parser and Linguistic Resources. Best Paper Award (2nd place). In *Proceedings of CICLing 2009*, Mexico City. Springer LNC 5449: 406-417.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam: John Benjamins.